IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jeffrey Toth
4/24/23

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection through Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Pandas and Matplotlib
  - Interactive Visual Analytics and Dashboards
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive Analytics
  - Predictive Analytics results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal of the project is to build a machine learning pipeline that can predict how successful the first stage of the launch will land.
- Problems you want to find answers
  - What factors determine if the rocket launch will land successfully?
  - How do various interactions among features affect the success rate?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Performed one-hot encoding to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data was collected using various methods
  - A get request was used on the SpaceX API to collect the data
  - Next we decoded the response content using the .json() function and converted it into a pandas dataframe using json_normalize()
  - The data was filtered to only include the Falcon 9 launches
  - The missing Payload Mass values were replaced using .mean()
  - Additionally web scraping was performed using BeautifulSoup on Wikipedia for the Falcon 9 launch records
  - We exported the results to a csv file

Data Collection – SpaceX API

- We used a get request on the SpaceX API to access the data
- The data was decoded as a JSON and turned into a Pandas dataframe
- We filtered the data to only include the Falcon 9 launches, and did some basic data wrangling
- We exported the results to a csv file
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/Data%20Collection%20with%20API.ipynb

Data Collection - Scraping

- We performed web scraping on the “List of Falcon 9 Heavy Launches” Wikipage
- We started by using an HTTP get request on the Falcon 9 Launch HTML page, and created a BeautifulSoup object
- Next we extracted all column/variable names from the HTML table header
- We created a data frame by parsing the launch HTML tables
- Finally we exported the results to a csv file
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- We performed EDA to find patterns and determine any training labels
- We calculated the number of launches on each site, the number and occurrence of each orbit, and the number and occurrence of mission outcome per orbit type
- We created a binary landing outcome label from the outcome column
- The results were finally exported to a csv file
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- We created scatterplots to observe the relationship of Flight Number vs. Payload Mass, Flight Number vs. Launch Site, and Payload Mass vs. Launch Site for the Falcon 9 launches
- We created a bar chart to show and compare the success rate of each Orbit by class
- We created more scatterplots observe the relationship between Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type
- A line plot was created to visualize the Yearly average launch success rate
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a database and queried the data using SQL magic to find the following:
  - The names of the unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have a payload mass between 4000 and 6000
  - The total number of successful and failure mission outcomes
  - The names of the booster versions which have carried the maximum payload mass
  - The records which display the month names, failure landing outcomes in drone drop ship, booster version, and launch sites for months in the year 2015
  - A ranking of the counts of successful landing outcomes between 06/04/2010 and 03/20/2017 in descending order
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We used various map objects to represent different aspects of the Falcon 9 launches
    - We used one-hot encoding to assign launch failures to class 0, and launch successes to class 1
    - Markers were created to represent the different launch sites for the Falcon 9 launches
    - Circle clusters to represent the number of successes/failures at each launch site
        - Using color-labeled markers allowed us to identify which launch sites have a relatively success rate
    - Lines were used to represent the distance a launch site is from the nearest railroad, highway, and city
        - The distance was calculated by using the latitudes and longitudes of the launch site and specified landmarks
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

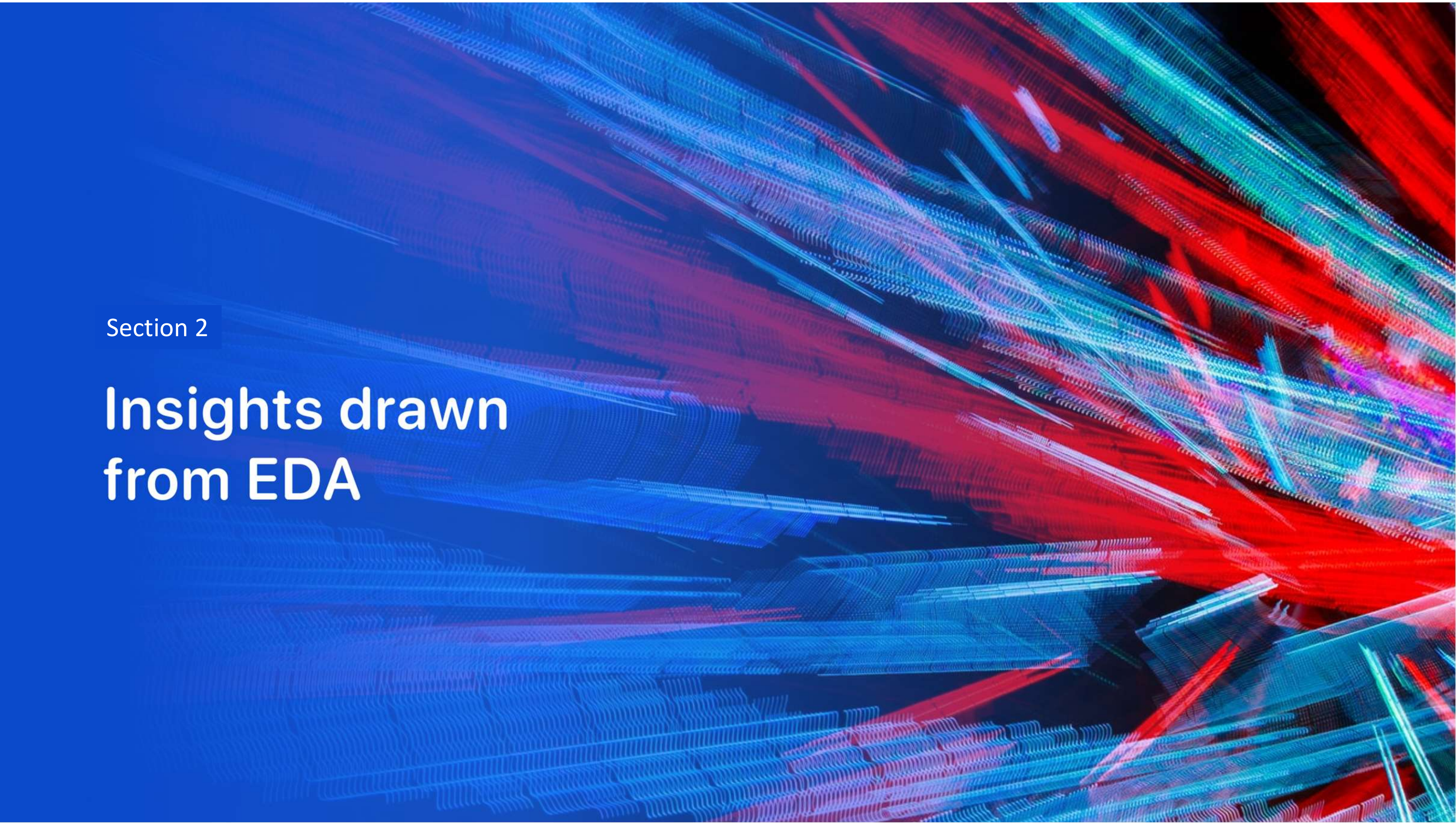
- We built an interactive dashboard using Plotly dash
- We plotted pie charts showing a breakdown of the total launches of all of the sites as well as the number of successes/failures of the specified launch site
- We plotted scatter plots to show the relationships between Payload Mass (kg) vs. Outcomes by different Booster Version Categories
  - A slider was also included to allow the user to adjust the Payload Mass to the desired amount
- GitHub URL for code for dashboard: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We created a NumPy array from the Class column in our dataset
- We standardized the data and assigned it to the variable X
- We split the model into a training and test set using the train_test_split function, with the test size set to 0.2 and the random state set to 2
- We used Logistic Regression, Support Vector Machine, Decision Tree, and K nearest neighbor machine learning models to fit our data and tuned different hyperparameters using GridSearchCV
- We found the accuracy of our validation data using the best_score_ attribute, and the accuracy of our test data using the score attribute for each of our models.
- We created a table comparing the accuracies of each model to find the best performing model
- GitHub URL: https://github.com/Jeff-Toth/IBM_Data_Science_Capstone/blob/main/Machine%20Learning%20Predictions.ipynb

Results

- Exploratory data analysis results
  - Over time the success rate of the launches increased
  - Orbits ES-L1, GEO, HEO, and SSO have a success rate of 100%, while SO has a success rate of 0%
- Interactive analytics
  - All launch sites are in close proximity to a body of water near the coast
  - The launch sites are far enough away from the nearest road, railway, and city, to where a failed launch shouldn't have an impact, but also close enough to have easy access if support is needed
- Predictive analysis results
  - Decision Tree model was our best predictive model for the dataset

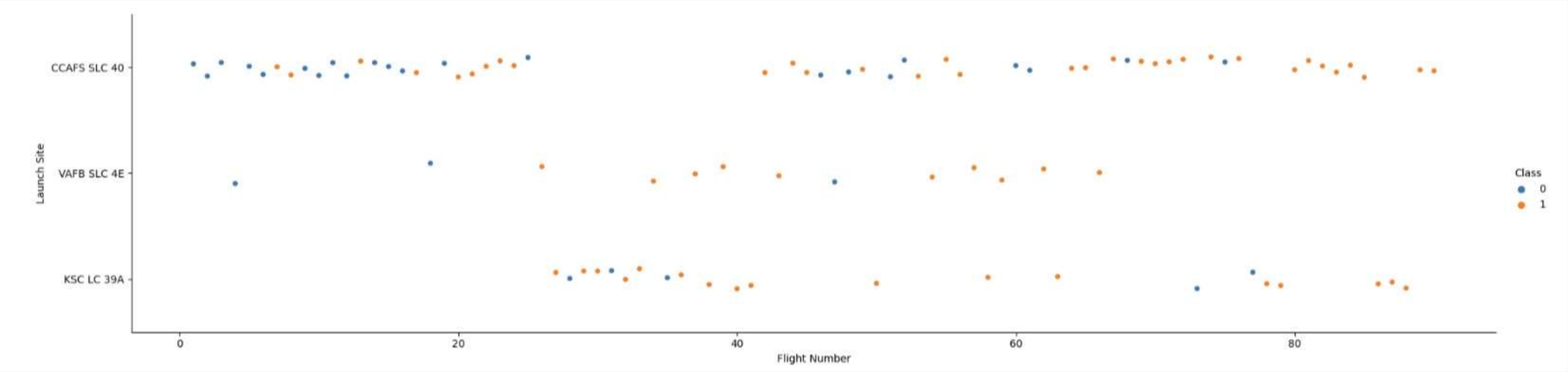


Section 2

Insights drawn from EDA

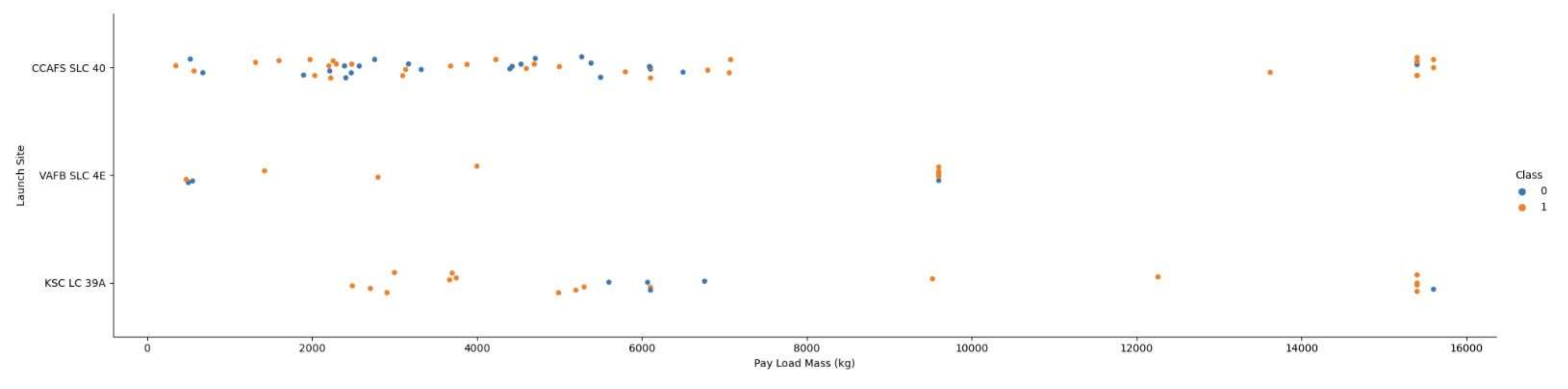
Flight Number vs. Launch Site

- Class 0 represents failed launches
- Class 1 represents successful launches
- It can be inferred that over time newer launches tended to have a higher rate of success compared to earlier launches with a smaller flight number



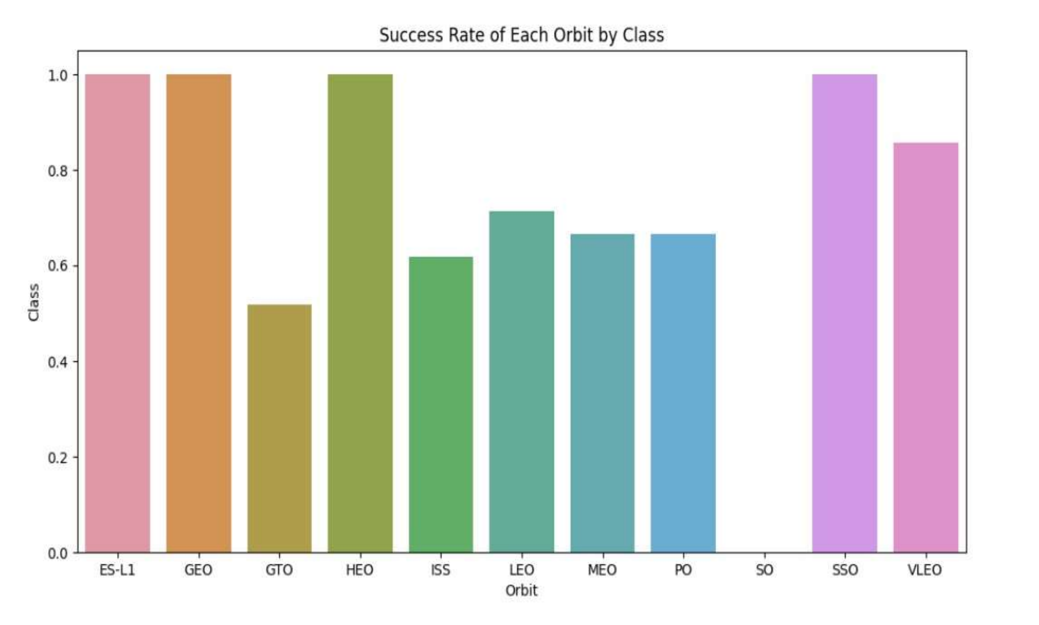
Payload vs. Launch Site

- Launches with a payload mass greater than 8,000 kg tended to be more successful than launches that had a payload mass less than 8,000 kg
- All launches with a payload ass less than 5,000 kg were successful from the KSC LC 39A launch site
- The VAFB SLC 4E has had no launches with a payload mass greater than 10,000 kg



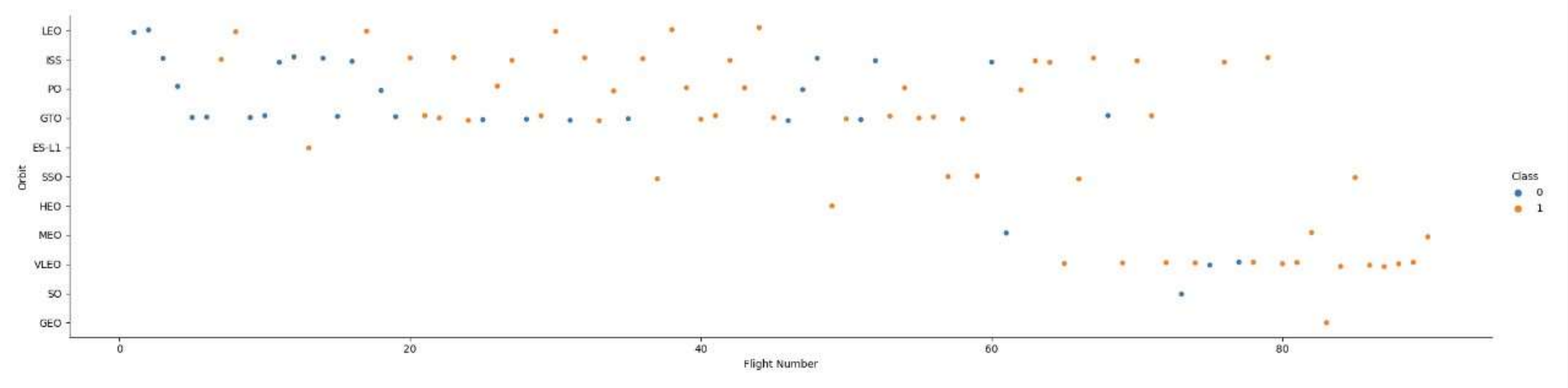
Success Rate vs. Orbit Type

- Each bar represents the average success rate for its perspective launch site
- ES-L1, GEO, HEO, and SSO had an average success rate of 100%
- SO had an average success rate of 0%
- GTO, ISS, LEO, MEO, PO, and VLEO had an average success rate between 50-85%



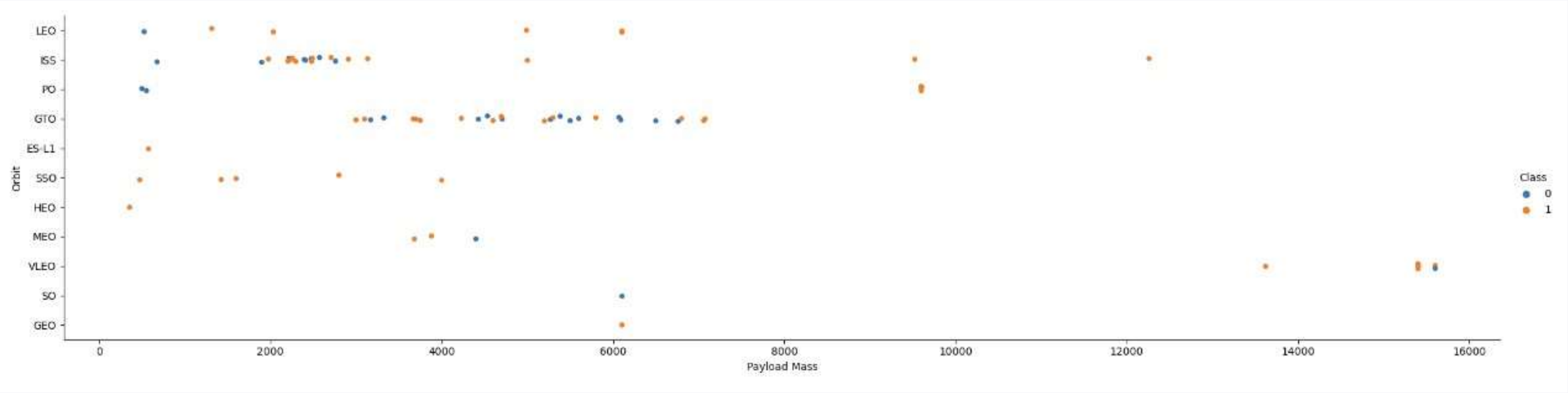
Flight Number vs. Orbit Type

- Orbits ES-L1, HEO, SO, and GEO only had 1 flight respectively
- The success rate typically increased as the number of flights increased for each orbit
  - Orbit GTO did not follow this trend



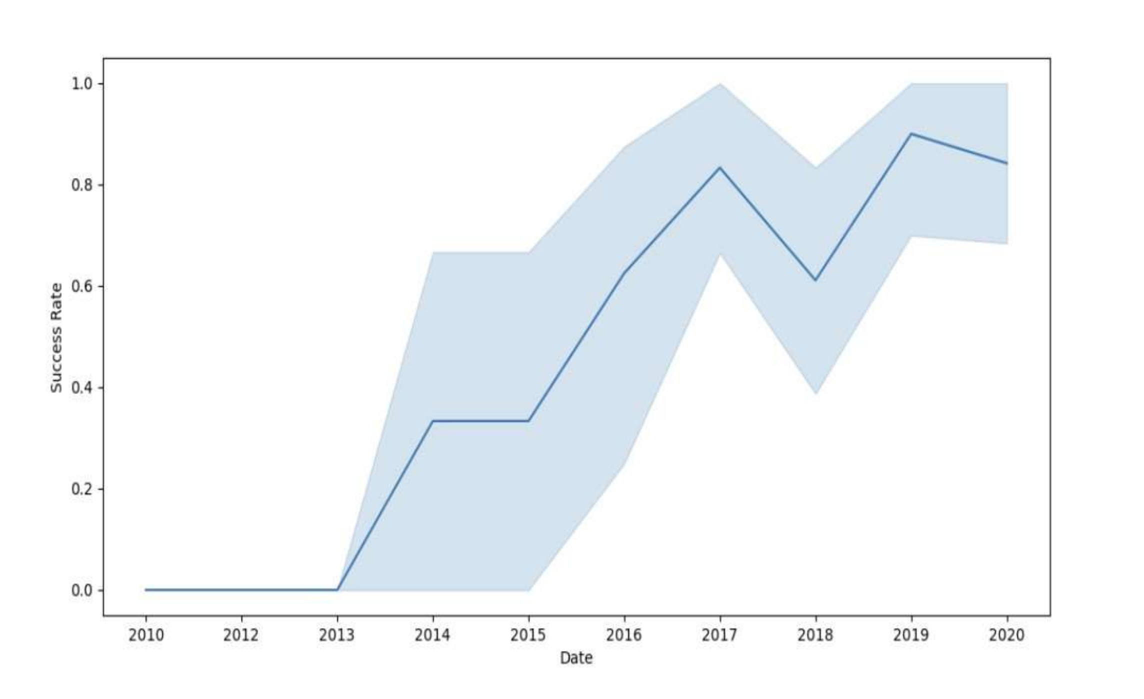
Payload vs. Orbit Type

- There were a much larger volume of payload masses less than 8,000 kg than payload masses greater than 8,000 kg
- Orbit VLEO was the orbit that contained the flights with the highest payload masses



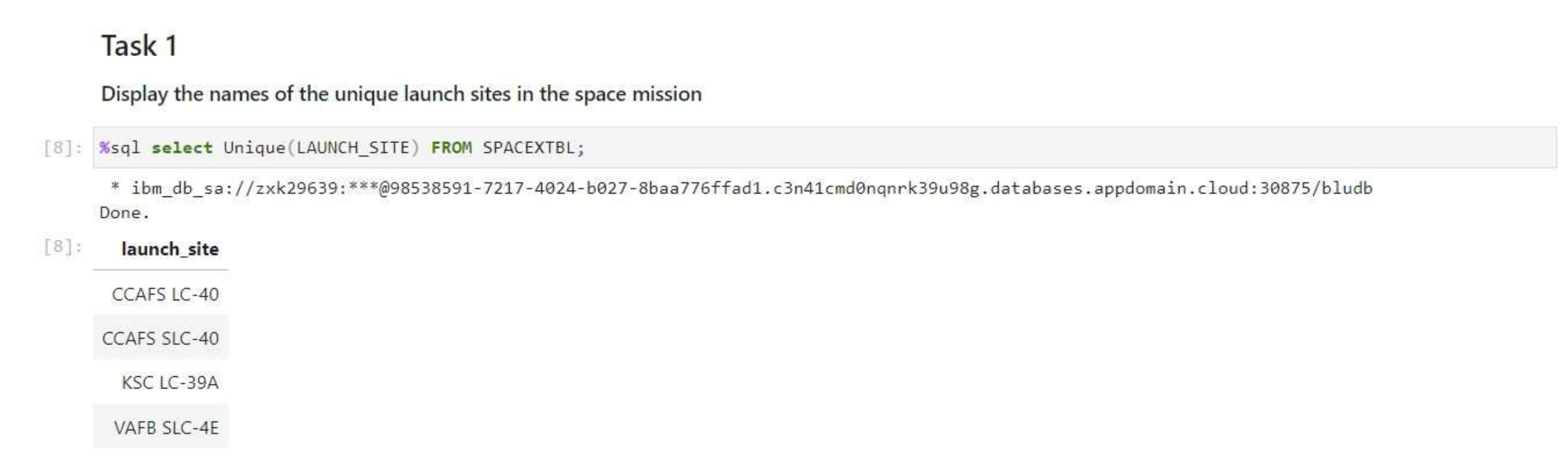
Launch Success Yearly Trend

- The success rate increased from 2013-2017, and from 2018-2019
- The success rate decreased from 2017-2018, and from 2019-2020
- The overall success rate has improved from 2013



All Launch Site Names

- We found the names of each unique launch site from the SpaceX database



Task 1

Display the names of the unique launch sites in the space mission

```
[8]: %sql select Unique(LAUNCH_SITE) FROM SPACEXTBL;
```

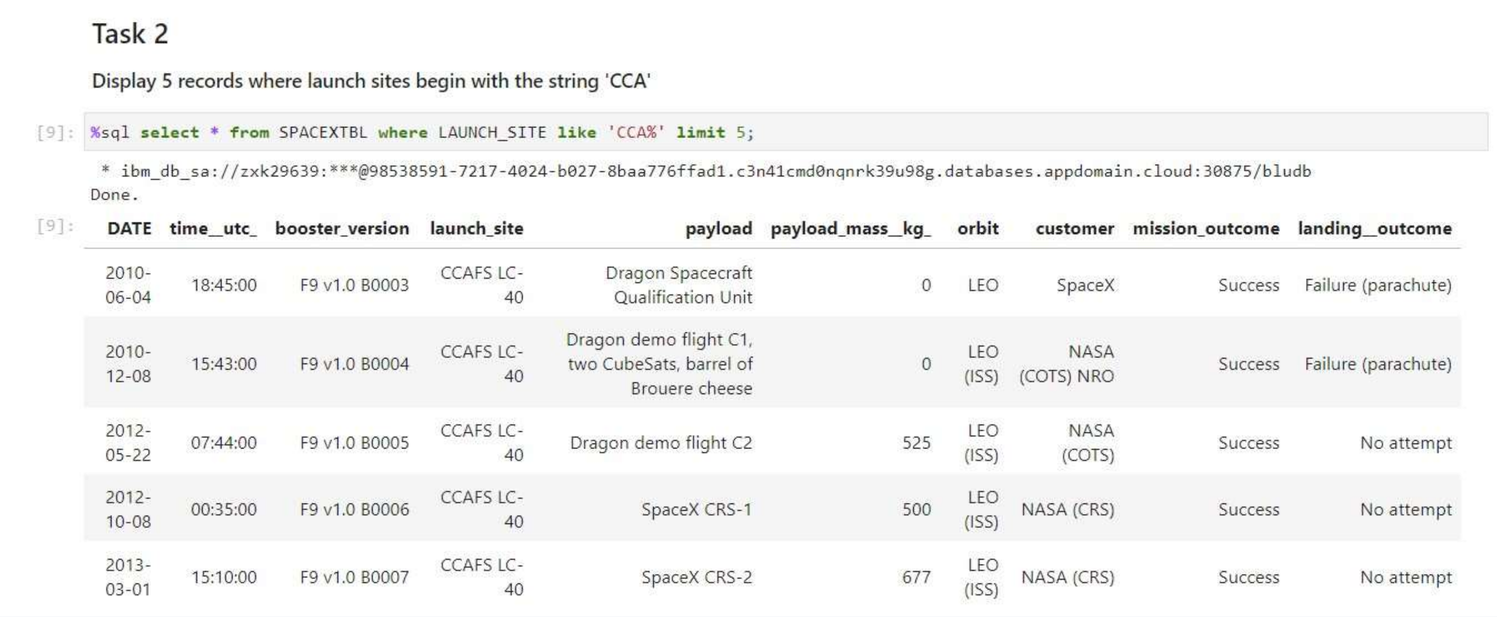
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[8]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- We found 5 records where launch sites began with ‘CCA’



Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
[9]: %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

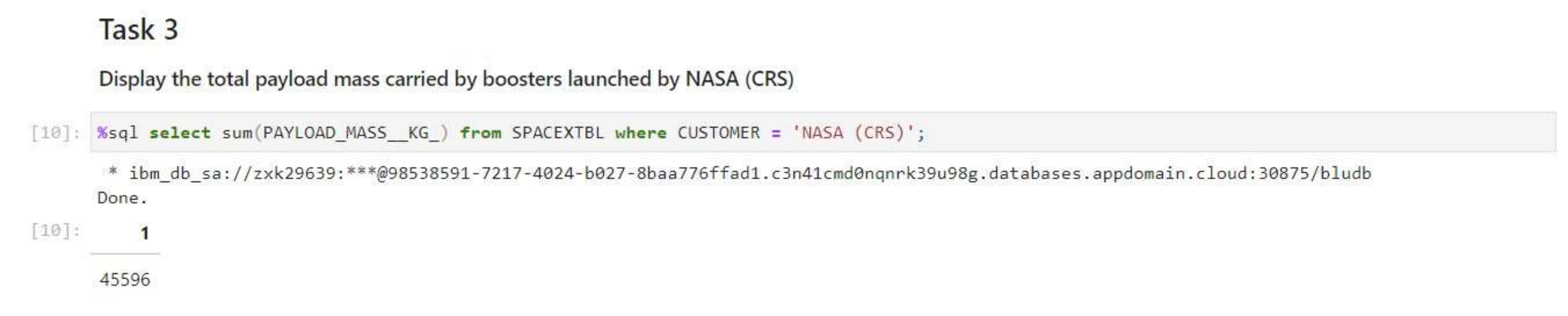
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[9]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We calculated the total payload carried by boosters from NASA which summed up to be 45,596 kg



Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

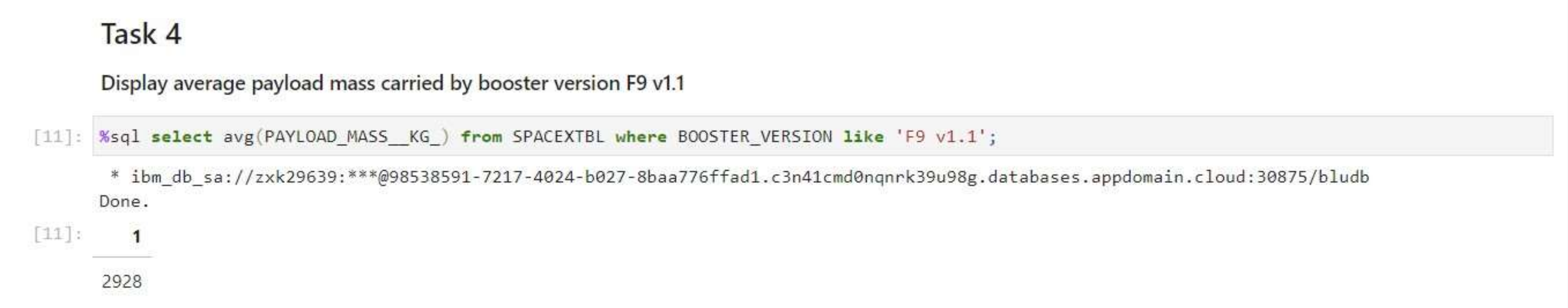
```
[10]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

```
[10]:     1
      -----
      45596
```

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 which averaged to 2,928 kg



Task 4

Display average payload mass carried by booster version F9 v1.1

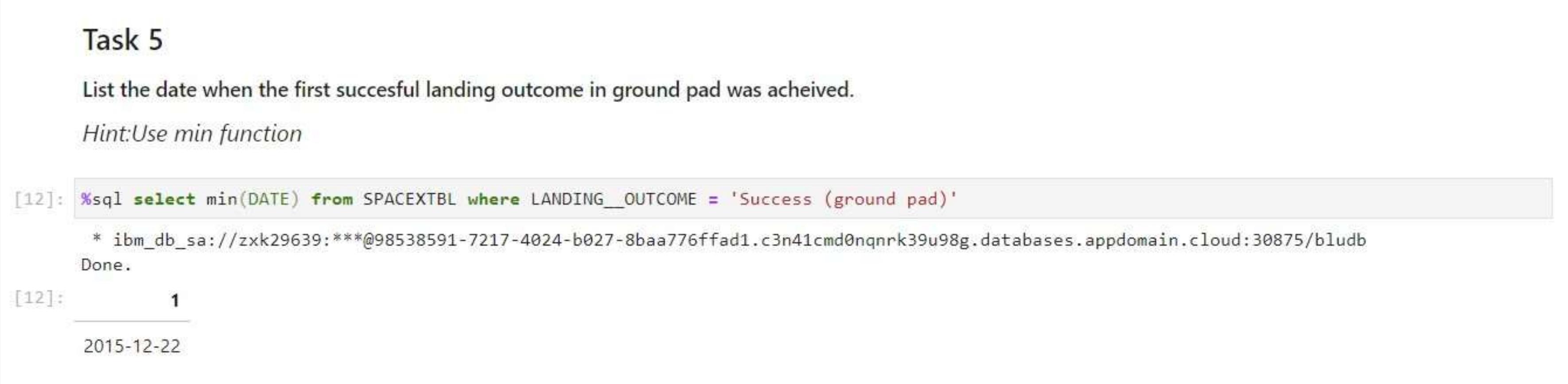
```
[11]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION like 'F9 v1.1';

 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.

[11]:   1
     2928
```

First Successful Ground Landing Date

- We found the date of the first successful landing outcome on ground pad which was 12/22/2015



Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
[12]: %sql select min(DATE) from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)'
```

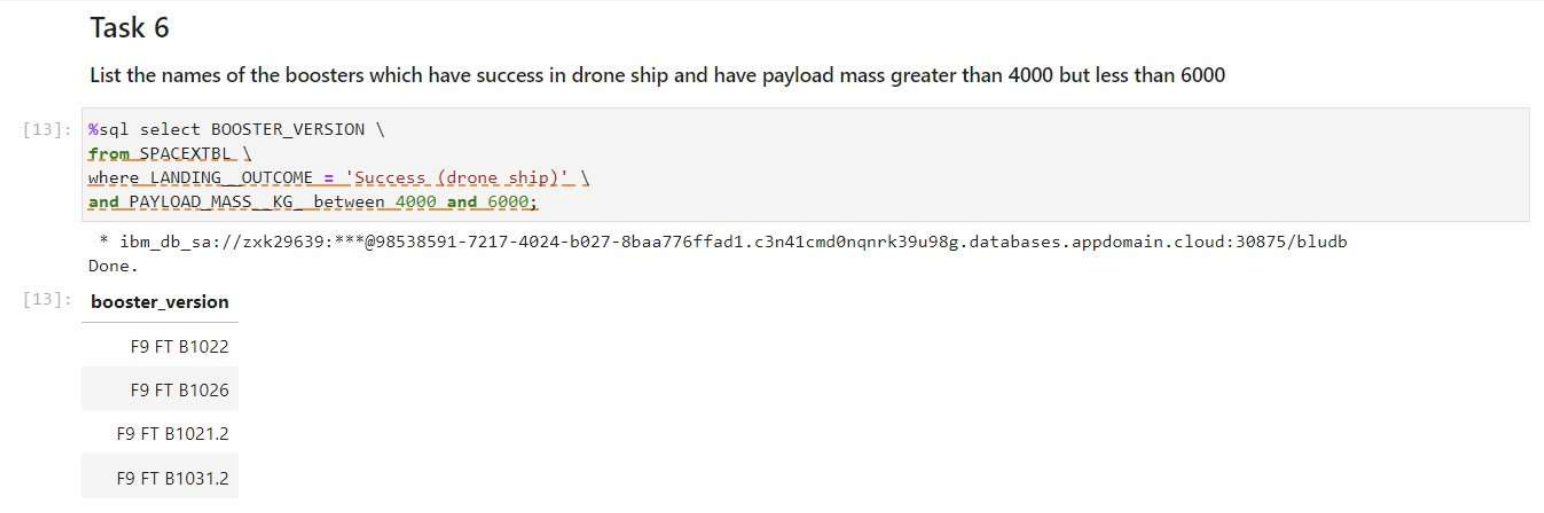
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[12]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We listed the names of boosters which have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000.
- These include: F9 FT B1022, F9 FT B1026, F9 FT V1021.2, and F9 FT B1031.2



Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[13]: %sql select BOOSTER_VERSION \
      from SPACEXTBL \
      where LANDING__OUTCOME = 'Success (drone ship)' \
      and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

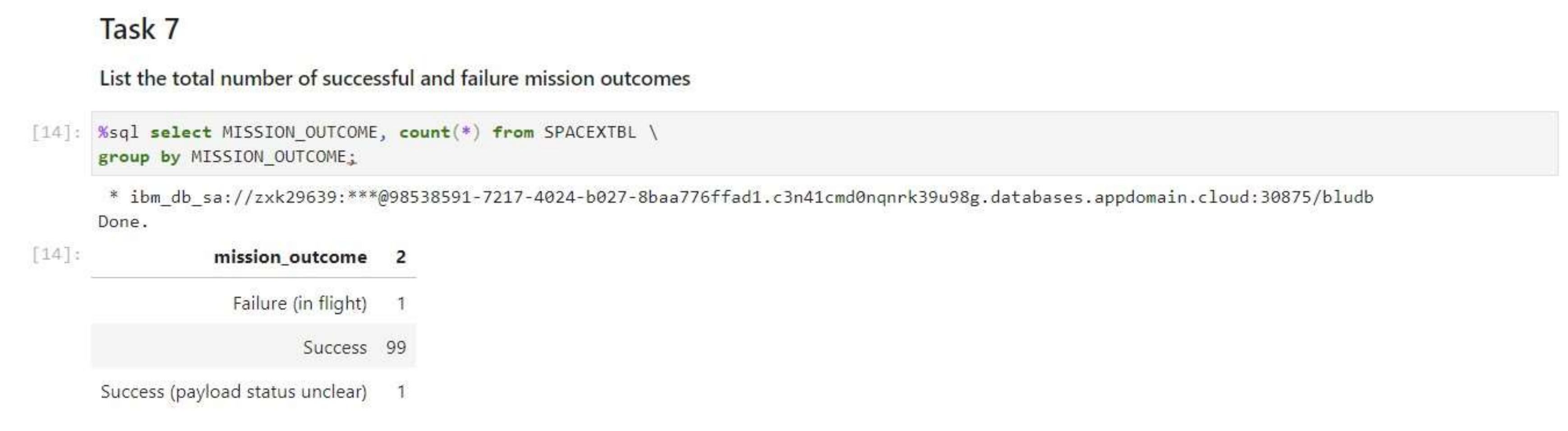
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[13]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- We found the total number of successful and failure mission outcomes which were:
  - 1 failure (in flight)
  - 99 successes
  - 1 success (payload status was unclear)



Task 7

List the total number of successful and failure mission outcomes

```
[14]: %sql select MISSION_OUTCOME, count(*) from SPACEXTBL \
      group by MISSION_OUTCOME;
```

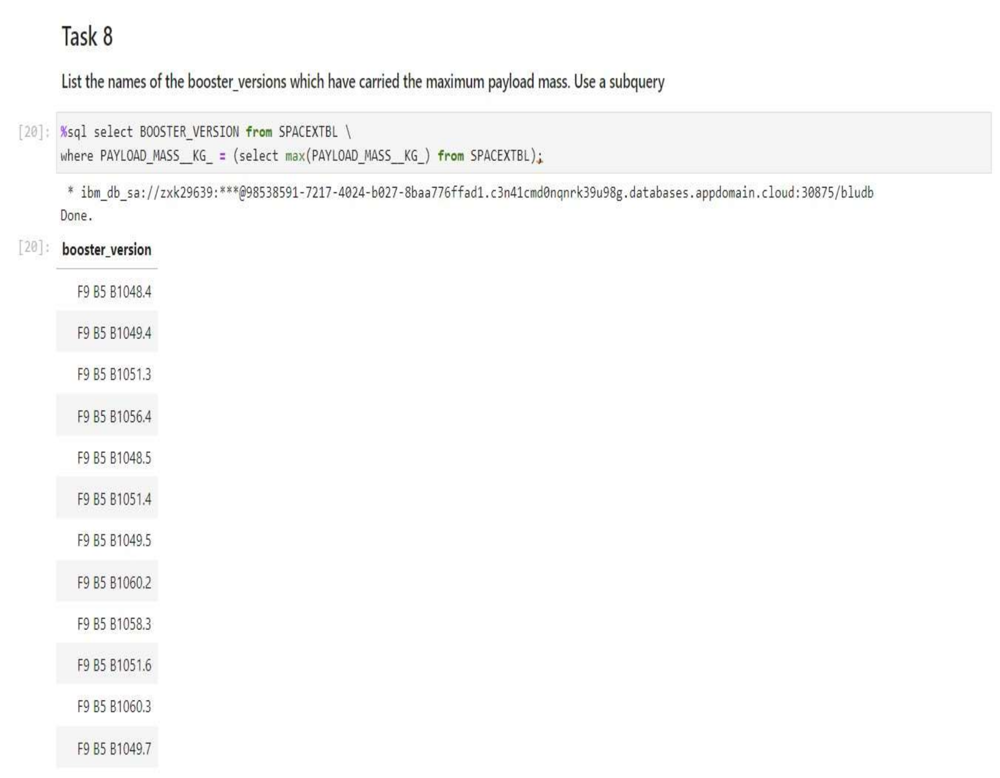
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[14]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- We listed the names of the booster which have carried the maximum payload mass which were:
  - F9 B5 B1048.4
  - F9 B5 B1049.4
  - F9 B5 B1051.3
  - F9 B5 B1056.4
  - F9 B5 B1048.5
  - F9 B5 B1051.4
  - F9 B5 B1049.5
  - F9 B5 B1060.2
  - F9 B5 B1058.3
  - F9 B5 B1051.6
  - F9 B5 B1060.3
  - F9 B5 B1049.7



Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[20]: %sql select BOOSTER_VERSION from SPACEXTBL \
      where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

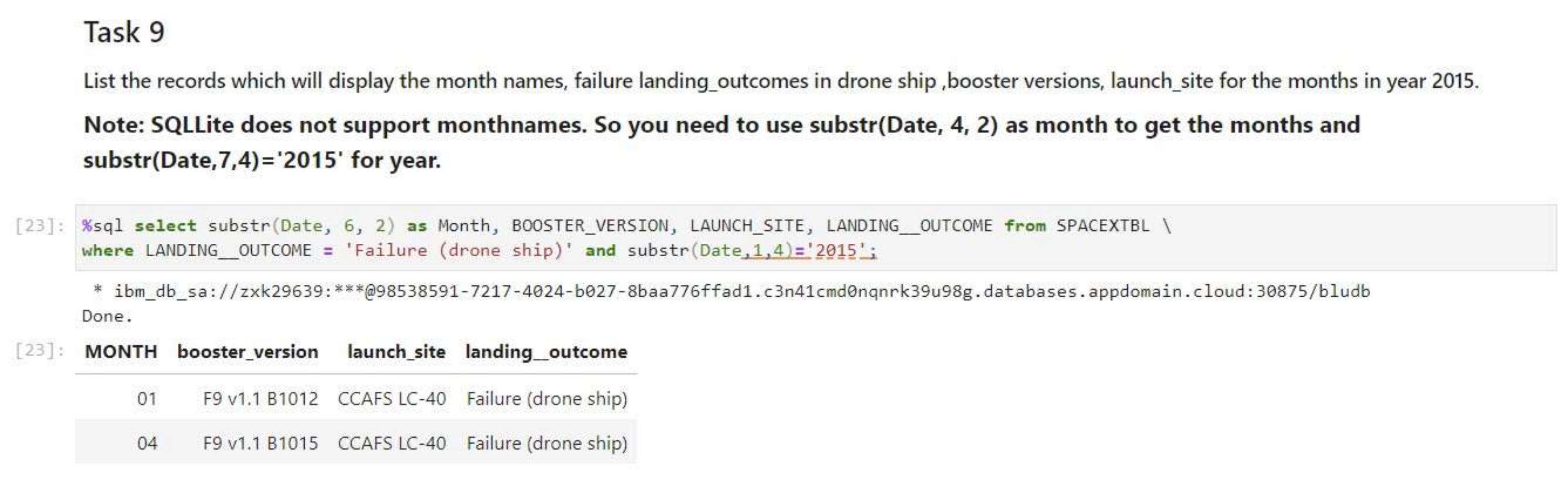
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[20]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- WE listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015



Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
[23]: %sql select substr(Date, 6, 2) as Month, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME from SPACEXTBL \
      where LANDING__OUTCOME = 'Failure (drone ship)' and substr(Date,1,4)='2015';
```

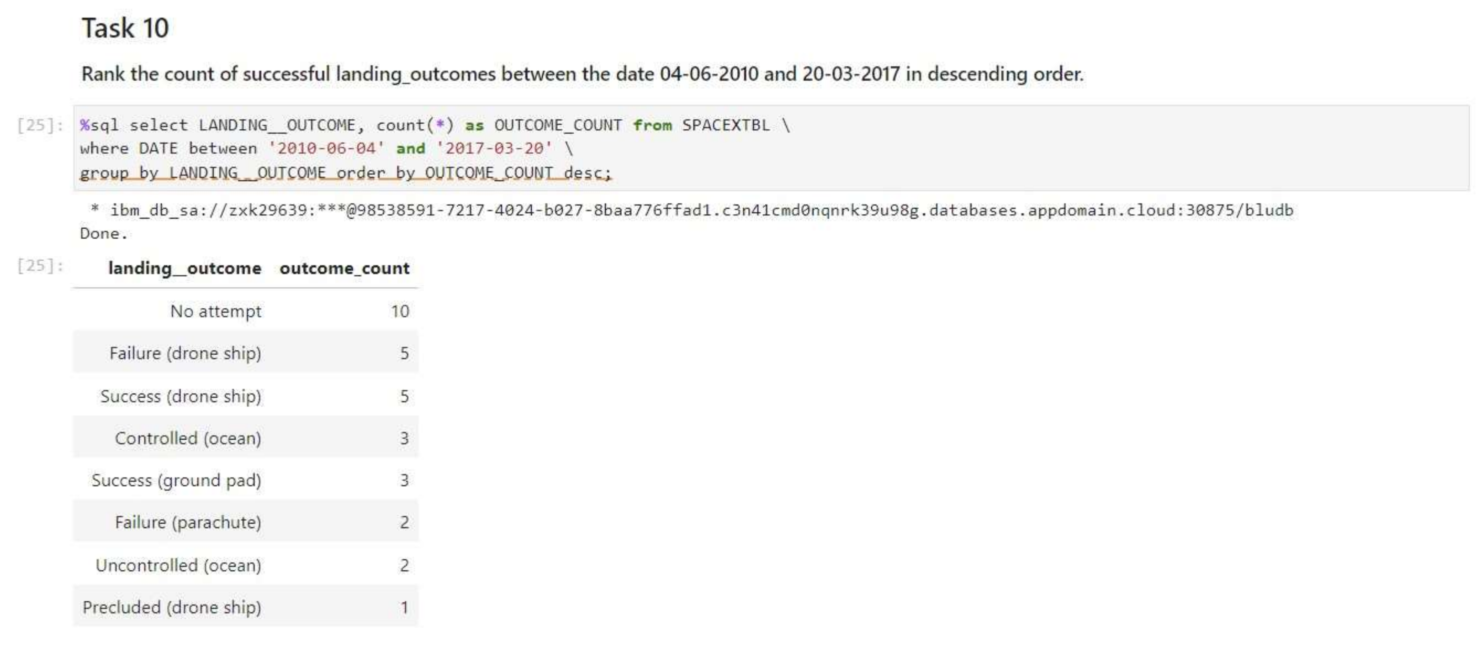
```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[23]:

MONTH	booster_version	launch_site	landing__outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- We ranked the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order



Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[25]: %sql select LANDING__OUTCOME, count(*) as OUTCOME_COUNT from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20' \
group by LANDING__OUTCOME order by OUTCOME_COUNT desc;
```

```
 * ibm_db_sa://zxk29639:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/bludb
Done.
```

[25]:

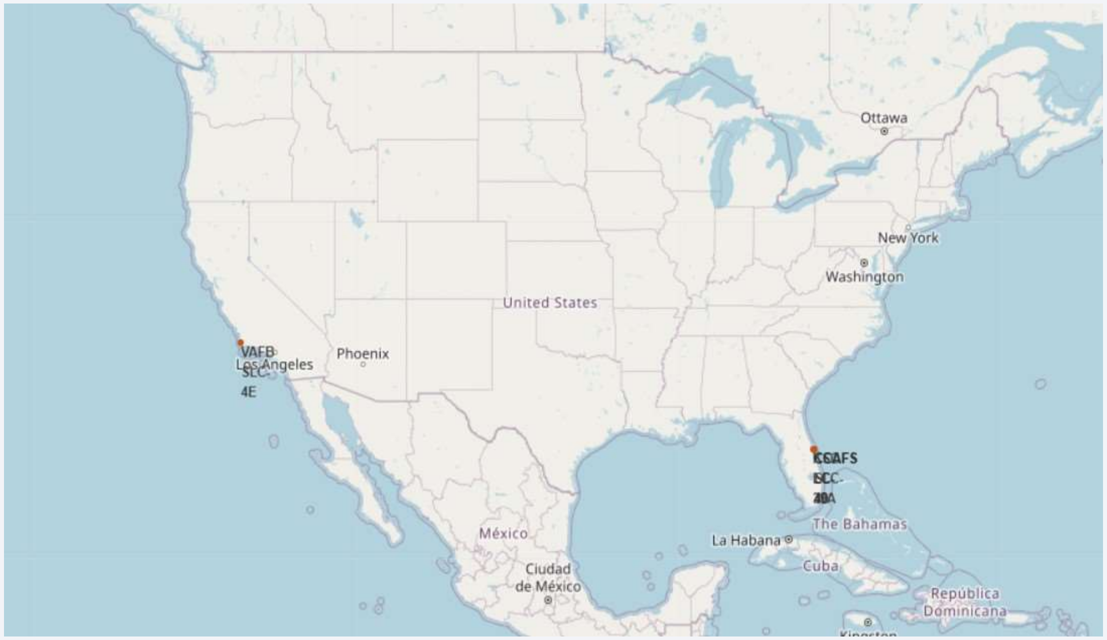
landing__outcome	outcome_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

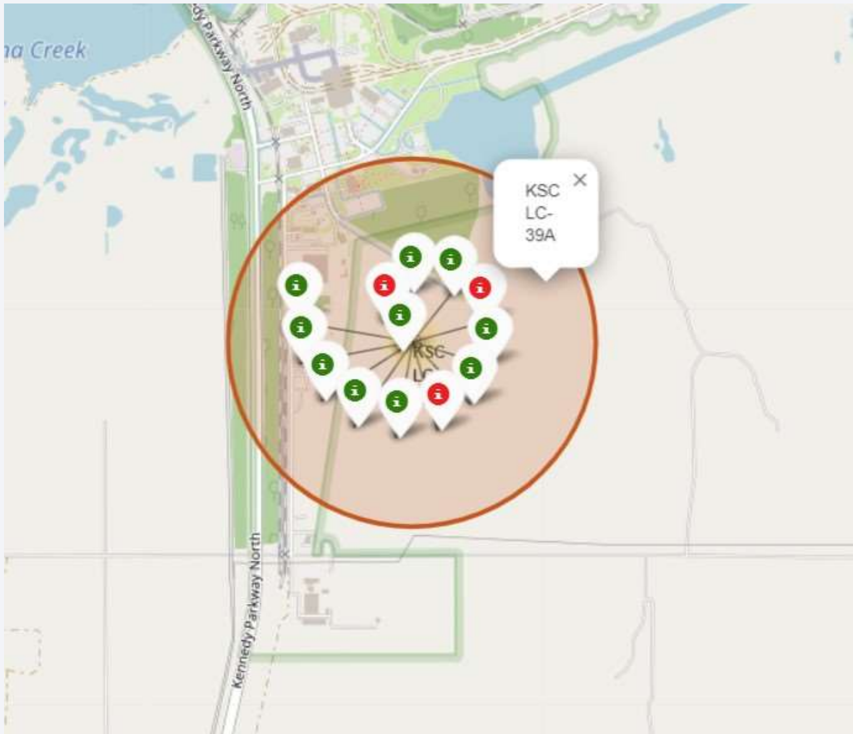
Launch Sites with Markers

- We can see that all launch sites are located on coasts in the United States
- The launch sites on the east coast are located in Florida
- The launch sites on the west coast are located in California

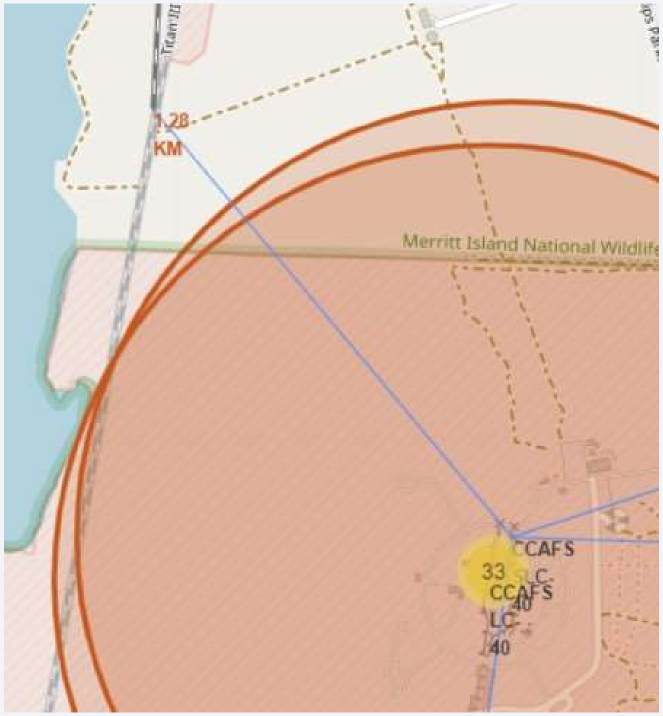


Launch Site with Color Labels

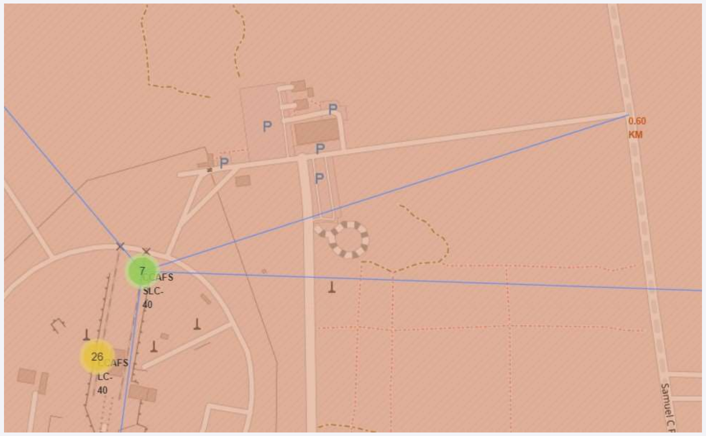
- For each site we can see the total number of launches, and if the launches were successful or a failure
- The map shows the launches at the site KSC LC-39A
  - Green markers represent successful launches
  - Red markers represent failed launches

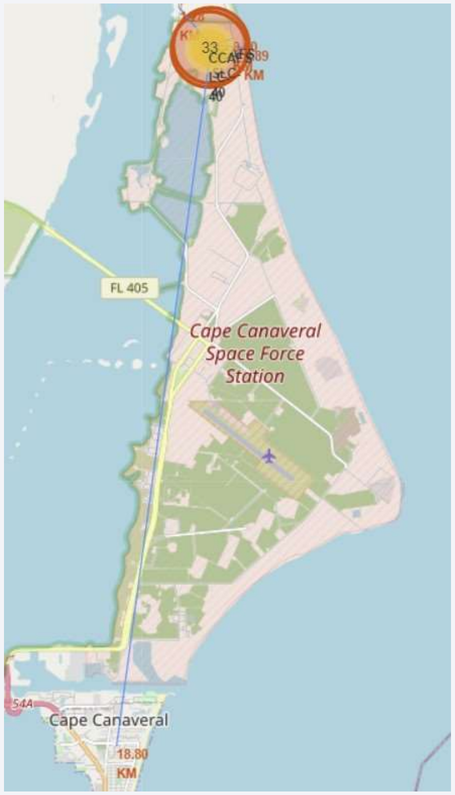


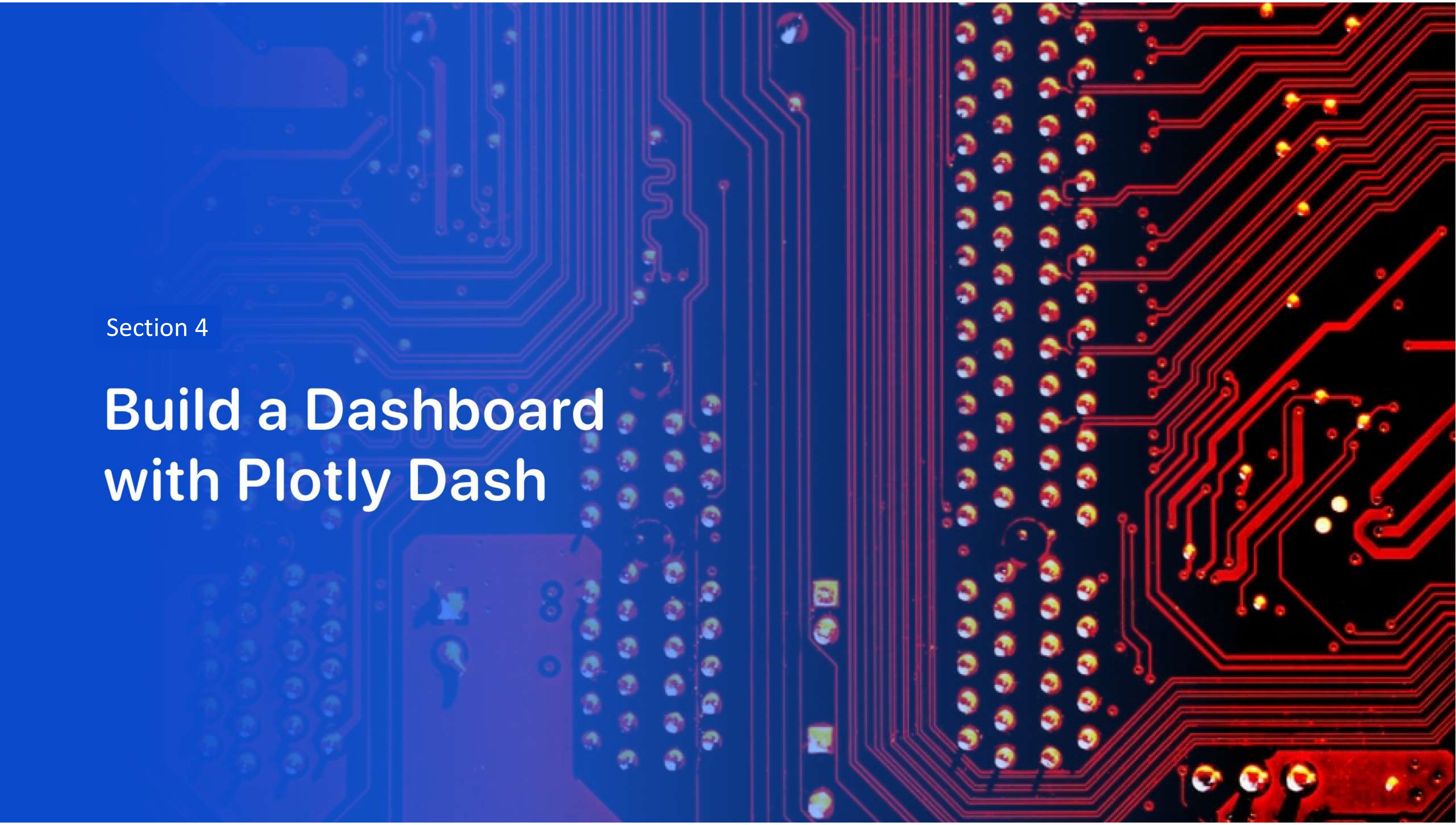
Launch Site Distance Markers



- We found the distance the CCAFS SCL-40 launch site is from the nearest railway, highway, and city
- The distance from each looks close enough to have support if needed, but far enough away to not cause any damage in the case of malfunctions or failed launches





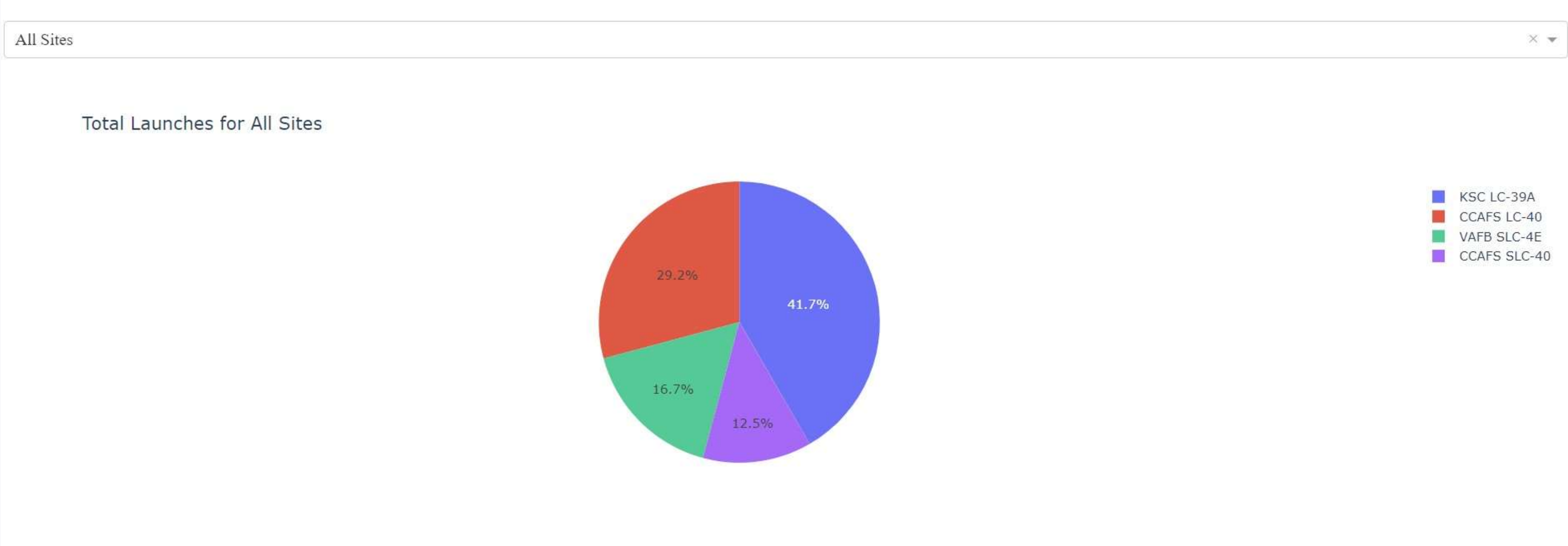


Section 4

Build a Dashboard with Plotly Dash

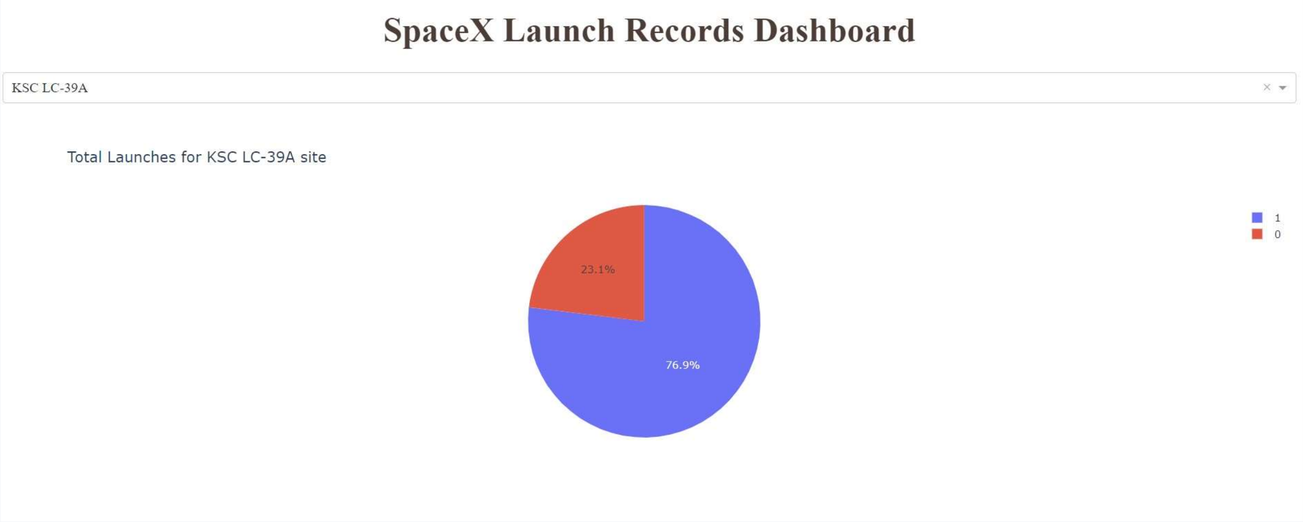
Number of Launches at Each Launch Site Pie Chart

- We can see that the largest number of launches occurred at the KSC LC-39A launch site, followed by CCAFS LC-40, VAFB SLC-4E, then CCAFS SLC-40 respectively



Highest Successful Launch Ratio Pie Chart

- The highest success ratio occurred at the KSC LC-39A launch site with 76.9% of launches being successful
  - Class 0 represents failed launches
  - Class 1 represents successful launches



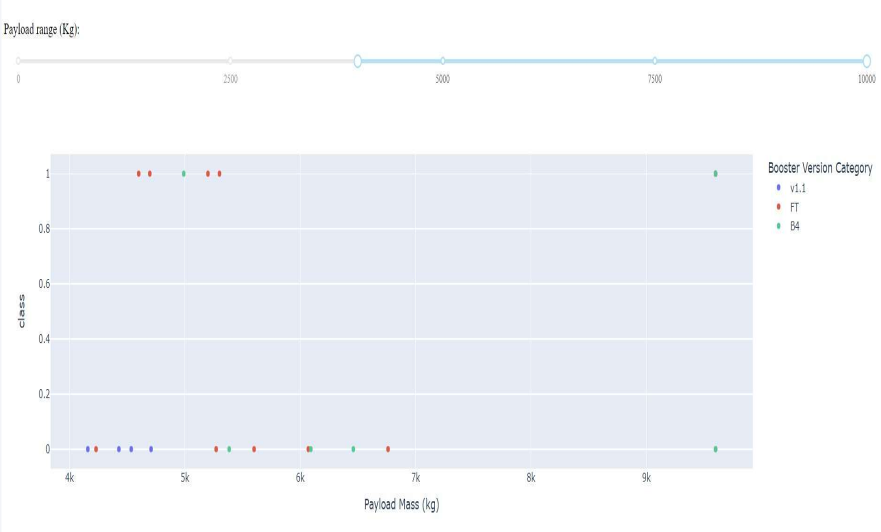
SpaceX Launch Records Dashboard

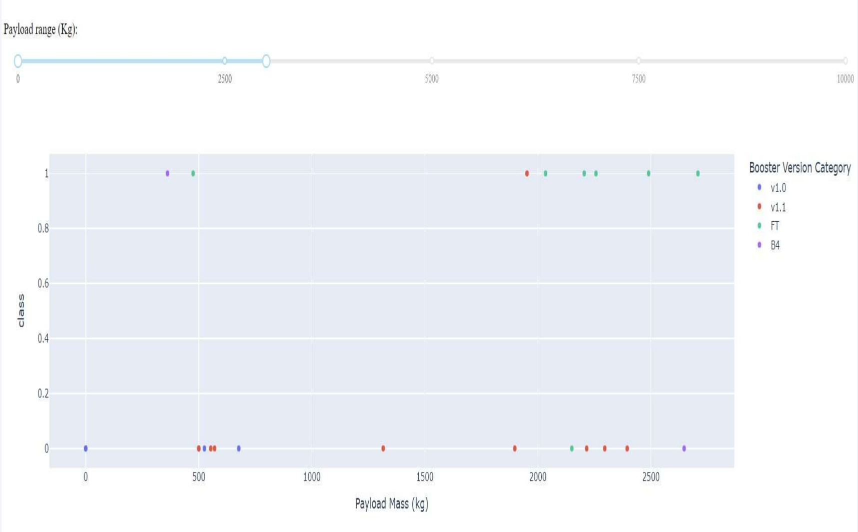
KSC LC-39A

Total Launches for KSC LC-39A site

Payload vs. Launch Outcome Scatterplot with Slider

- We can see that there were a larger number of successful missions when we selected a smaller payload range
- Out of all of the boosters, the FT Booster Version Category has the most successful outcomes in terms of volume and percentage



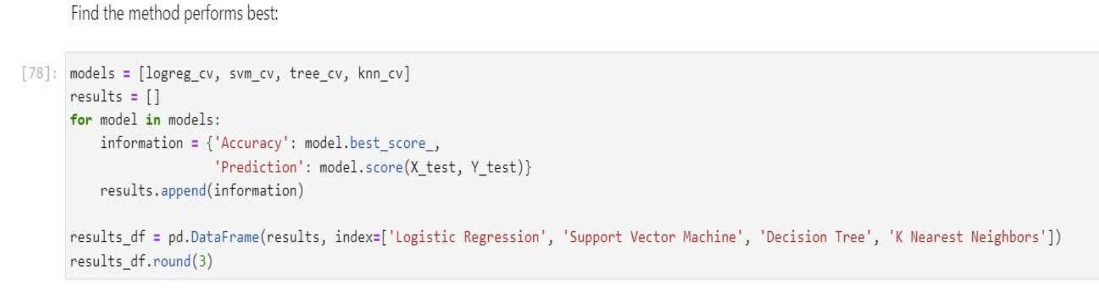


Section 5

Predictive Analysis (Classification)

Classification Accuracy

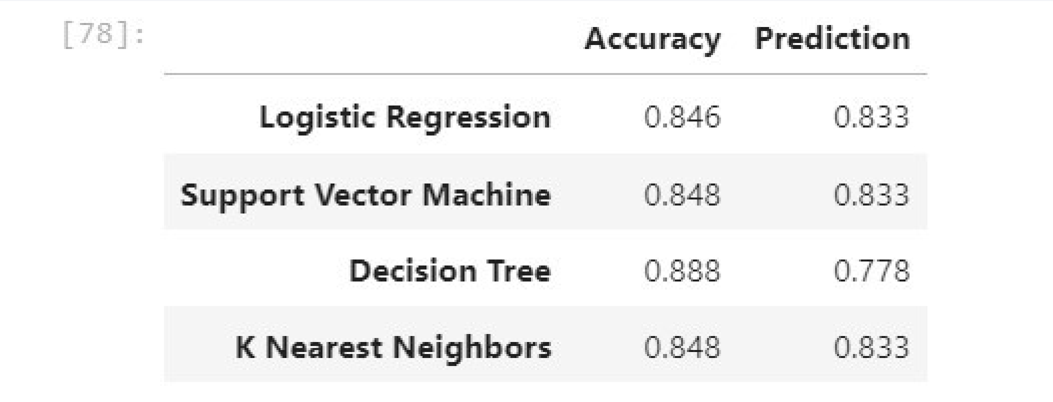
- Of the 4 models we used, the decision tree classifier was the model with the highest classification accuracy based off .best_score_ function



Find the method performs best:

```
[78]: models = [logreg_cv, svm_cv, tree_cv, knn_cv]
      results = []
      for model in models:
          information = {'Accuracy': model.best_score_,
                         'Prediction': model.score(X_test, Y_test)}
          results.append(information)

      results_df = pd.DataFrame(results, index=['Logistic Regression', 'Support Vector Machine', 'Decision Tree', 'K Nearest Neighbors'])
      results_df.round(3)
```

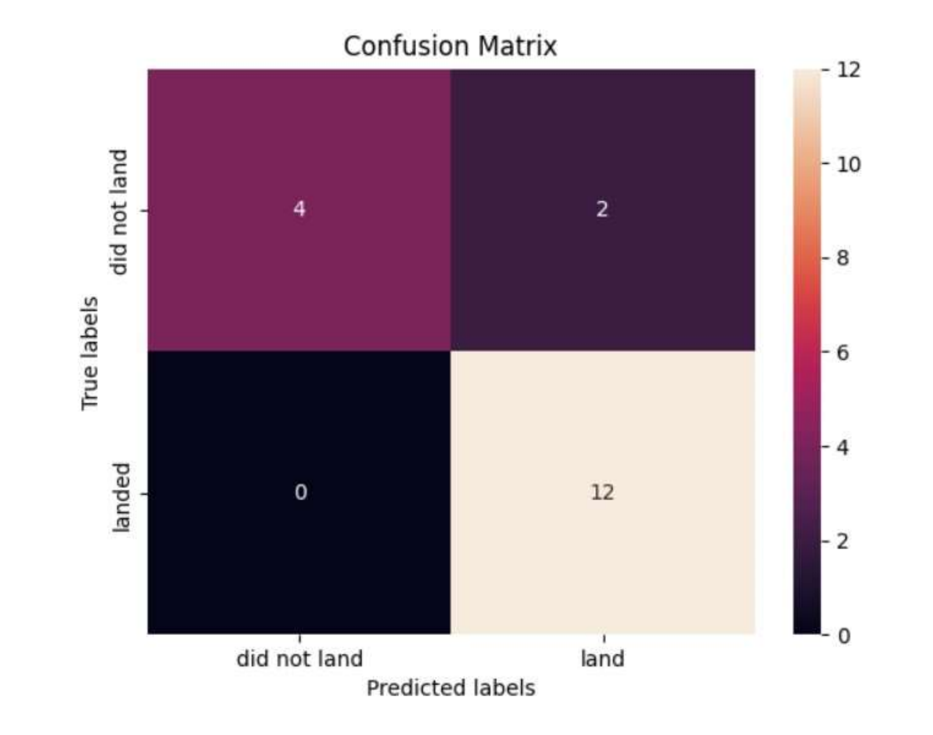


[78]:

	Accuracy	Prediction
Logistic Regression	0.846	0.833
Support Vector Machine	0.848	0.833
Decision Tree	0.888	0.778
K Nearest Neighbors	0.848	0.833

Confusion Matrix

- The confusion matrix outputs for our decision tree classifier model are displayed
- Our results show the following:
  - 4 True Negatives
  - 2 False Positives
  - 0 False Negatives
  - 12 True Positives
- Having False positives is not good as it labels unsuccessful landing as successful ones by the classifier



Conclusions

- The launch success rate saw an overall increase over time from 2013-2020
- The KSC LC-39A launch site was the site that had the highest percentage of successful launches
- ES-L1, GEO, HEO, and SSO orbits had the highest success rate of 100%
- Our decision tree classifier was the classification model that yielded the highest accuracy out of our ones tested

Thank you!